DOCUMENT RESUME

ED 463 322 TM 033 754

AUTHOR Wheelock, Anne
TITLE School Awards Programs and Accountability in Massachusetts: Misusing MCAS Scores To Assess School Quality.
PUB DATE 2002-03-20
NOTE 46p.
PUB TYPE Opinion Papers (120)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Accountability; *Achievement Gains; Achievement Tests; *Educational Quality; Elementary Secondary Education; High Stakes Tests; Rewards; State Programs; *Test Results; Testing Programs
IDENTIFIERS *Massachusetts; Massachusetts Comprehensive Assessment System

ABSTRACT

This paper makes the case that the high-stakes testing and accountability program currently in place in Massachusetts misuses scores from the Massachusetts Comprehensive Assessment System (MCAS) to select particular schools as exemplary. School awards and recognition programs in Massachusetts imply that score gains are equivalent to school progress, but, in fact, score increases do not necessarily signify either improved student learning or school quality. Rising MCAS scores are poor signposts to best practices for replication by other schools, and scores may even benefit from policies and practices that harm or neglect the most vulnerable students. In many Massachusetts schools listed as exemplary, the statistical patterns associated with small numbers of students tested, changes in the composition of a schools students taking the MCAS from one year to the next, and teaching to the test may artificially improve test scores without improving school quality. By using MCAS score gains to identify particular schools as models of school improvement, public policymakers and pro-MCAS corporate leaders promote an inadequate definition school quality, misrepresent schools cited for test scores as more exemplary than others, and do a disservice to parents, teachers, and students who seek authentic school improvement and who care more about public education than public relations. (Contains 118 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 463 322

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

A. Wheelock

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

School Awards Programs and Accountability in Massachusetts: Misusing MCAS Scores To Assess School Quality

Anne Wheelock
18 Cranston Street
Jamaica Plain, MA 02130
Phone: 617.524.7324
Email: wheelock@shore.net

March 20, 2002

TM033754

BEST COPY AVAILABLE

School Awards Programs and Accountability in Massachusetts: Misusing MCAS Scores To Assess School Quality

The School and District Accountability System is the shining star of education reform in that it's taking schools for what they are, where they're starting off, and allowing them to show what they can do in terms of improvement. - MA Department of Education spokesman Jonathan E. Palumbo, quoted in Tantraphol, 2001.

Last year we had the second best MCAS scores in the state, yet, according to the DOE, we have two `failing' schools. We don't believe we are above reproach. There is certainly room for improvement in virtually everything we are trying to do here. But if the Department of Education is trying to embarrass people into doing better on these tests, I'm not sure that it is going to work. Hopefully, people are going to be smart enough to say, the emperor has no clothes on. - Gary Burton, Superintendent, Wayland Public Schools, quoted in Caruso, 2001.

The question is, are we picking out lucky schools or good schools, and unlucky schools or bad schools? The answer is, we're picking out lucky and unlucky schools. - David Grissmer, RAND Corporation, quoted in Olson, 2001.

Since 1998, the release of scores from the Massachusetts Comprehensive Assessment System (MCAS) has become an annual event anticipated by journalists, business groups, parents, educators, and real estate brokers. When scores decline, policy makers and the media call on students and teachers to "try harder," while district leaders attempt to pinpoint reasons for lack of progress. When scores rise, educators celebrate, and policy makers and local media point to the evidence that reforms are working.

When it comes to raising MCAS scores, the stakes are high. Schools that make gains stand to win financial awards from several public and private school recognition programs. And entire communities look to scores to establish their ranking in relation to other communities (Pappano, 2001). But do MCAS score gains accurately indicate school improvement? Are higher test scores conclusive signs that school quality is improving? Can scores pinpoint which schools should be recognized as exemplary or which practices deserve replication?

A review of Massachusetts MCAS-based awards programs challenges the assumption that MCAS score gains are accurate and appropriate measures of school improvement. The high-stakes testing and accountability system currently in place in Massachusetts *misuses* MCAS scores to select particular schools as "exemplary." And although schools awards and recognition programs imply that score gains are tantamount to school progress, in fact, score increases do not necessarily signify either improved student learning or school quality. Rising MCAS scores, moreover, are poor signposts to "best practices" for replication by other schools. To the contrary, scores may even benefit from policies and practices that harm or neglect the most vulnerable students.

How can test score gains in award-winning schools be inaccurate measure of school qualtiy?

- Many schools cited as "exemplary" on the basis of short-term score gains do not sustain gains for all four years of MCAS testing. In most award-winning schools, the percentages of students scoring at combined "advanced/proficient" and "failing" levels bounce up and down from one year to the next.

- In many schools cited as "exemplary," the number of students tested is so small that MCAS score gains may have more to do with luck and statistical patterns than with authentic improvement in learning. When numbers tested are small, especially in schools testing 68 students or less, the presence of a few "stars" or "slow" students can change scores dramatically from year to year, making score gains or drops unreliable indicators of school quality.

- Score gains in some schools likely reflect changes in the composition of students taking MCAS rather than any instructional improvement. Scores may increase because of differences in student characteristics from one cohort to another, or, in the case of high schools, because grade retention in ninth grade or tenth grade attrition may remove weaker students from the testing pool.

- Increases in students leaving school earlier in the high school grades can push up tenth grade MCAS scores. In the majority of award-winning high schools and vocational schools recognized for MCAS score gains from one year to the next, 2000 dropout rates were higher than in 1997, the year before MCAS testing started.

- Widespread reports of teaching to the test in tandem with data that suggest diminishing holding power for weaker students suggest that schools may be more focused on producing higher test scores in order to *look* good than on making improvements in teaching and learning that result in authentically better schooling for all students.

This paper examines a number of flaws in the test-based approach to Massachusetts school recognition programs specifically and the state's use of test scores for school accountability. It also describes how a more authentic accountability system holds greater promise for building school capacity, improving student learning, and strengthening school holding power than do current policies. Grounded in the assumption that reform depends on a positive partnership between local districts and the state, alternative approaches would develop resources and strengthen professional accountability for equity, excellence, and high standards of teaching and learning for all students without high stakes.

School awards programs in Massachusetts

> *We hope the... awards will serve as an incentive for all principals to strive to facilitate real change in their schools.* - [Then Lt.] Gov. Jane Swift, quoted in Perlman, 2000.

School awards programs in Massachusetts mimic similar test-based accountability practices in other states. Policy makers typically establish school recognition programs in the belief that "rewards for performance" will motivate professionals to work above and beyond normal levels of effort to improve test outcomes in their schools (Blair, 1998; Bradley, 1996; Walsh, 2000). These recognition programs use test scores to rank schools, elevate particular schools deemed high-performing to "exemplary" status, and herald practices in these schools as worthy of adoption in others.

In Massachusetts, three programs use MCAS scores to identify "most improved" or "exemplary" schools. First, since 1999, the privately-funded Edgerly School Leadership Awards program, founded by William Edgerly, Chairman Emeritus of the State Street Corporation and

Chairman of the Foundation for Partnerships, has annually recognized five to 10 principals from schools deemed to have made the greatest MCAS score gains from one year to the next and where at least 40 students were tested. Each recognized principal receives $10,000 to be used at his discretion (Massachusetts Department of Education, 1999; Massachusetts Department of Education, 2000; Massachusetts Department of Education, 2001b).

Second, MassInsight Corporation, comprised of business groups and "partner" school districts, has used MCAS scores to name 12 Massachusetts schools, including eight elementary schools, one middle school, and three high schools, and one district as "2001 Vanguard Schools." Selected on the basis of score gains and an application process that asks schools to describe how they have developed high-standards curricula, strengthened teaching, used data to improve learning, and intervened to help struggling students, award-winning schools are recognized at a high-profile MassInsight-sponsored conference and in publications that describe practices for other schools to emulate (Gehring, 2001; also see, MassInsight, Building Blocks Initiative Publications: http://www.massinsight.com/meri/Building%20Blocks/e_bb_press.htm).

Third, beginning in the school year 2000-2001, the Massachusetts Department of Education released the first of its biannual school performance ratings based on annual MCAS scores. Using scores from the first round of testing in 1998 as a baseline, Massachusetts places schools in a "performance category" ranging from "1" (for top-scoring schools) to "6" (for low-scoring schools) for both overall and content-specific performance. The Department then sets school-specific expectations for score gains, requiring schools at the lowest levels to make the largest score improvements. Schools receive two ratings: a performance rating based on the average of the 1999 and 2000 MCAS results and an improvement rating based on the comparison of those results to the 1998 baseline results, with schools cited as "failing to meet,"

"approaching," "meeting," and "failing to meet" expectations (Massachusetts Department of Education, NDa).

From the outset, observers have cited factors ranging from mathematical errors to the strong correlation between MCAS scores and community income to argue that the Massachusetts school rating system is misconceived and misleading (Bolon, 2001; Caruso, 2001; Haney, 2002; McElhenny, 2001b; Moore, 2001; Sutner & McFarlane, 2001; Tantraphol, 2001; Tomei, 2001; Tuerck, 2001; Vaishnav, 2001). Despite widespread criticism, the state's Department of Education publishes a list of schools that have met or exceeded MCAS improvement expectations and invites schools listed to apply to become a "Commonwealth Compass School" (Massachusetts Department of Education, NDc). Based on this ranking process, the Department has named 14 schools, including 10 elementary and 4 middle schools as 2001 Compass Schools. These schools also receive $10,000 each in the expectation that they will "promote improvement in student performance by sharing their experiences with other schools in the state" (Massachusetts Department of Education, NDb, Massachusetts Department of Education, 20 December, 2001).

Test score gains: A poor measure of school quality

> *We happened to do very well the first year. One elementary school was in the top 1 percent of the state. It turns out you'd have been better off doing really bad the first year.* - Peter Manoogian, director of curriculum and technology, Lynnfield Public Schools, Hayward, 2001a).

Together, the three Massachusetts awards programs portray MCAS score gains as a fair and accurate means of assessing school quality. But drawing conclusions about school quality or the merit of particular practices on the basis of score gains is risky at best, duplicitous at worst. In fact, MCAS scores in award-winning schools selected in the early rounds of citations typically do not show steady improvement over four years of testing. Small numbers of students tested in

many schools, changes in the composition of test-takers, and widespread teaching to the test can all influence MCAS scores, including in schools cited as "exemplary," without making appreciable improvements in authentic student achievement.

Test score ups and downs: Observations from other states

Over the past decade, a number of states have forged ahead with programs that reward or sanction schools based on test scores, even in the fact of research that suggests the limitations of test scores for drawing conclusions about either school or instructional quality (see, for example, Bauer, 2000; Stecher & Barron, 1999). Close observers of state testing programs have long noted that test scores patterns are predictable, typically rising in the early years of testing, then leveling off and declining over time as scores regress to the mean (Camilli & Bulkley, 2001; Darling-Hammond, 1997; Hoff, 2000; Koretz, Linn, Dunbar, & Shepard, 1991). Not surprisingly, then, the record of school accountability programs in states that routinely grade schools on the basis of test scores underscores how test scores, including test score *gains*, are imprecise measures of school improvement.

Policy assumptions to the contrary, short-term test score gains appear to be especially poor predictors of score increases in subsequent years. In Florida, where the state's school grading program demands annual improvement, schools rated "A" one year regularly rate "C" the next (Palm Beach Post, 2001). In North Carolina and Texas, wide swings in schools' test scores have been so common that over the past decade, virtually every school in those states could have been categorized "failing" at least once (Kane & Staiger, 2001; Kane, Staiger, & Geppert, 2001).

Since 1994, Kentucky's department of education has used scores from its annual state tests to classify schools biannually into particular categories (formerly "Rewards," "Successful," "In decline," and "In crisis," now "Meets goals," "Progressing," and "Needs assistance"). The

approach has grown more disruptive over time as schools cited in the top category one year have found themselves ranked in the bottom category two years later (Darling-Hammond, 1997; Frommeyer, 1999; Whitford & Jones, 2000). In Pennsylvania, many award-winning schools have also failed to sustain gains following an initial burst of progress. Of 85 Philadelphia schools recognized for improvements on the state test in 1999, only 15 also made gains in that qualified them for awards in 2000. Indeed, most 1999 award-winning Philadelphia schools produced score declines in 2000, including 29 of 36 schools that won awards for eighth grade score gains (Socolar, 2001). Commenting on these patterns, Philadelphia Public School Notebook editor Paul Socolar says, "For anyone paying any attention to this stuff, it's obvious that we're celebrating a different group of 'high performing schools' each year" (Personal communication, 7/23/01).

MCAS score swings in award-winning schools

Over time, MCAS gains in award-winning or "exemplary" schools will likely to prove as unstable as score gains from other states. Score patterns of schools winning the first two rounds of Edgerly School Leadership Awards illustrate the extent to which MCAS scores are unreliable measures of school quality, with schools selected as "most improved" in early citations showing drops in scores in later years. (This paper draws on Massachusetts Department of Education, November 21, 2000, and Massachusetts Department of Education, November 2001a, for data on MCAS scores and participation rates).

Notably, MCAS gains in Edgerly award-winning schools in 1998 and 1999 have not set the stage for continued gains in 2000 and 2001. Not one of the five early award-winners steadily increased the rate of students scoring in "advanced" or "proficient" categories in both English and math in each of the four years of testing while also steadily reducing the percentage of students

scoring "failing" in these subjects. Rather, between 1998 and 2001, results were erratic, showing little consistency.

- At the Franklin D. Roosevelt in Boston, testing an average of 54 fourth-graders annually, scores have bounced up and down dramatically from 1998 through 2001. In English, the percentage of students scoring "advanced" or "proficient" went from 0% in 1998 to 20% in 1999, dropping back to 9% and 4% in 2000 and 2001 respectively. In math, the percentage of students in these top score categories went from 2% in 1998 to 53% in 1999, then back to 22% in 2000 and 27% in 2001. Although the percentage of students "failing" English dropped from 33% in 1998 to 2% in 1999, and registered 7% in 2000, and 9% in 2001, the percentage of students "failing" in math was more volatile in these years, totaling 65%, 2%, 30%, and 18% in the years of testing.

- At Kensington Avenue School in Springfield, where an average 46 fourth-graders are tested annually, early score gains were not sustained over four years. After expanding the number of students scoring at proficient levels in English from 2% in 1998 to 40% in 1999, 0% scored at this level in 2000, bouncing back to 50% in 2001. Kensington's math scores proved equally volatile. In 1998, 12% of the students scored at the "advanced" or "proficient" levels in 1998, jumping to 70% in 1999, then dropping to 39% in 2000 and 43% in 2001. In 1998, 1999, 2000, and 2001, English "failing" rates were posted at 29%, 0%, 9%, and 6%, and math "failing" rates were posted at 17%, 2%. 7%, and 13% respectively.

- At the Abraham Lincoln in Revere, testing an average of 83 fourth-graders, 5% of students scored at "advanced" or "proficient" levels in English in 1998, rising to 27% in 1999, then dropping to 14% in 2000, and rising again to 44% in 2001. In math, 12% scored at "advanced" or "proficient" levels in 1998, rising to 38% in both 1999 and 2000, but dropping to 27% in 2001. During this period, "failing" rates in English bounced from 10% in 1998 to 1% in 1999, rising to 5% in 2000, and 8% in 2001. Likewise, "failing" rates in math fell from 43% in 1998 to 6% in 1999 and 2000, but rising again to 14% in 2001.

- At Riverside Elementary School in Danvers, testing an average of 58 fourth-graders, results were more favorable, but still mixed. In English, the percentage of students scoring at "advanced" or "proficient" levels has increased steadily from 9% in 1998 to 72% in 2001, while the percentage "failing" has dropped from 11% in 1998 to 2% in 2001. However, in math, "advanced" and "proficient" score levels have been more erratic, rising from 15% in 1998, to 54% in 1999 and 69% in 2000, but dropping again to 45% in 2001; "failing" rates have dropped from 21% in 1998 to 3%, 2%, and 2% in 1999, 2000, and 2001 respectively.

- At Swampscott High School, testing an average of 183 tenth graders, the percentage of students scoring at proficient and "advanced" levels increased in both English and math over four years. While the percentage of students "failing" in math decreased, the percentage of students "failing" English was more volatile, dropping from 24% 1998 to 14% in 1999, returning to 24% in 2000, then dropping to 5% in 2001.

Announcing the first round of Edgerly Awards, William Edgerly stated, "These principals have lead (*sic*) their schools to impressive improvement. By their example, they heighten appreciation of the principal's role, and direct attention toward future possibilities for our schools" (Massachusetts Department of Education, 1999).

However, despite policy-makers' optimism, "future possibilities" did not include continued improvement in MCAS scores from 1999 to 2001. Moreover, MCAS scores in schools recognized in Edgerly's second round of awards also showed sharp bounces over four years of testing. In several of the Round 2 Edgerly schools, gains made from 1999 to 2001 simply returned school score levels to those of 1998. In others, gains made from 1999 to 2000 were not sustained in 2001, or the number tested in 2001 was so small as to render "improvement" ambiguous. Specifically:

- At Hopkinton High School the percentage of students scoring at "advanced" or proficient levels in English dropped from 63% to 38% from 1998 to 1999, rising back to 67% in 2000 and up to 77% in 2001, while the percentage of students "failing" increased from 9% in 1998 to 26% in 1999, dropping back to 9% in 2000 and 2% in 2001. In math, the percentage of students scoring "advanced" or "proficient" totaled 44%, 32%, 68%, and 72% in the four years of testing, while "failing" rates bounced from 14% to 41%, then back to 15% and 6%.

- At Nantucket High School from 1998 to 1999, the percentage of students scoring at "advanced" and "proficient" levels in English had dropped while the percentage of "failing" scores had increased, making gains in 2000 and 2001 appear significant when, in fact, they simply returned to early levels. Thus, percentages scoring at "advanced" or "proficient" levels in English were 55%, 40%, 57%, and 59%, while percentages scoring at "failing" levels were 15%, 24%, 12%, and 14% in the years tested. In math, percentages posted for "advanced" or "proficient" levels were 38%, 16%, 69%, and 60%, while those for "failing" were 39%, 49%, 21%, and 15%.

- At Lowell Middlesex Academy Charter School, scores remained steady in 1998 and 1999, with 27% and 25% scoring "advanced" or "proficient" in English and 4% scoring "advanced" or "proficient" in math for both years; in those years 23% and 22% scored "failing" in English, and 85% and 82% scored "failing" in math. In 2000, the percentage scoring at "advanced" or "proficient" jumped to 51% in English, and 21% in math, while the "failing" rate dropped to 4% in English and 40% in math. Rates at high and low levels remained about the same in both

subjects in 2001, but that year only 16 students were tested that year, making these percentages virtually meaningless.

- At Seven Hills Charter School, the percentage of eighth graders scoring at "advanced" or "proficient" levels in English dropped from 40% in 1998 to 21% in 1999, then rose to 55% in 2000, dropping again to 51% in 2001; at the same time the percentage "failing" rose from 23% in 1998 to 33% in 1999, then dropped to 5% in 2000, rising slightly to 8% in 2001. In math, the percentage of students scoring at "advanced" or "proficient" levels dropped from 21% in 1998 to 5% in 1999 before rising again to 22% in 2000 and dropping again to 19% in 2001 while the percentage "failing" rose from 62% in 1998 to 80% in 1999, before falling again to 46% in 2000 and 50% in 2001.

- At Marshfield's Eames Way Elementary School, testing an average of 48 students, the percentage of students scoring "advanced" or "proficient" has moved steadily up, with 26%, 25%, 55%, and 96% at those levels in English, and 50%, 49%, 81%, and 85% scoring at these levels in math in 1998, 1999, 2000, and 2001 respectively. For four years, the percentage of students "failing" English has remained at 0%; that percentage "failing" math went from 2% to 16% between 1998 and 1999, dropping to 0% in 2000 and 2001.

Like MCAS scores from the Edgerly schools, scores from MassInsight's Vanguard Schools have also proved unstable over time. None of the 12 schools cited has shown steady four-year increases at "advanced" or "proficient" levels and steady declines in "failing" for both English and math. Two - Hudson High School and Longmeadow's Williams Middle School - have come close to doing so. However, other schools show notable bounces in either English or math scores over four years. For example:

- In Everett, although the Devens School shows a steady increase of students scoring at "advanced" or "proficient" levels and a steady low of only 2% of test-takers "failing" in English, the percentage of students at the "advanced" or "proficient" levels in math has dropped precipitously from 58% in 1998 to 28% in 2001, and the percentage "failing" math has risen from 2% in 1998 to 13% in 2001. Likewise, although Everett's Lewis School shows an increase in the percentage of students scoring "advanced" or "proficient" in English, rising from 0% 1998 to 26% in 2001, the percentages in math are not so decisively improving, with percentages of students at "advanced" or "proficient" moving from 15% to 17% between 1998 and 1999, to 64% in 2000, but dropping to 22% in 2001, and "failing" rates moving from 19% to 6% to 0% from 1998 to 2000, but back to 12% in 2001.

- In Woburn, the positive increase in the percentage of fourth graders scoring "advanced" or "proficient" in English (rising from 24% in 1998 to 65% in 2001) at the Altavesta School is offset by an increase in the percentage of students "failing" (rising from 3% in 1998 to 22% in 2001). Altavesta's math scores have bounced around considerably, with percentages at the

"advanced" or "proficient" level rising from 24% to 68% between 1998 and 1999, and again to 85% in 2000, but dropping to 42% in 2001, and the percentages "failing" posted at 18% in 1998, 0% in 1999 and 2000, and 25% in 2001. Likewise, gains in moving higher rates of students into "advanced" and "proficient" categories in math have not been sustained at either the Goodyear or Reeves Schools in Woburn, and the percentage of Goodyear students scoring at "advanced" or "proficient" levels in English has bounced from 61% to 75% from 1998 to 1999, back down to 55% in 2000, and up again to 81% in 2001.

- At Arlington's Thompson School, after the percentage of fourth graders scoring "advanced" or "proficient" in math rose from 54% in 1998 to 82% in 1999, "advanced/proficient" rates dropped back to 72% in 2000 and 64% in 2001.

Finally, the state's own Compass Schools show similar fluctuating score patterns over four years. Again, none of the 14 schools showed steady increases in percentages of students scoring at "advanced" or "proficient" levels and steady declines in "failing" rates in both English and math, and only four - Quincy's Sterling Middle School, East Somerville Middle School, Boston's - - Longmeadow's Williams Middle, and Orleans Elementary - came close. Otherwise, these schools, including Edgerly's Kensington Avenue and Riverside Schools, showed sharp score gains some years, declines the next, in either English or math, or both. For example:

- MCAS scores at Westfield's Moseley School did not sustain gains after 1998. Moseley's scores showed early increases in "advanced/proficient" levels, rising from 6% to 19% in English and from 6% to 27% in math from 1998 to 1999. However, in English, the percentages of students scoring "advanced/proficient" dropped back to 10% in 2000, then rising to 20% in 2001, while percentages of students "failing" rose to 23% in 2000, staying at 20% in 2001. In math, the percentages of students scoring "advanced" or "proficient" dropped from 27% back to 23% in 2000 and again to 11% in 2001, while the percentages of students "failing" declined to 19% in 2000, only to rise again to 38% in 2001.

- Although MCAS scores at Salem's Saltonstall School improved after the first year of testing, by 2001, the school's "failing" rates in both English and math equaled the rates for 1998. From 1998 to 1999, the percentage of Saltonstall's students scoring at "advanced" or "proficient" levels rose from 13% to 29% in English, and from 29% to 45% in math, while "failing" rates dropped from 16% to 12% in English, and from 28% to 18% in math. However, the percentage of students scoring "advanced" or "proficient" in English dropped back to 20% in 2000, before rising again to 55% in 2001, while the "failing" rate dropped to 9% in 2000, only to return to 17% in 2001. At the same time, while the percentage of students scoring "advanced" or "proficient" in math rose to 52% in 2000, it declined to 35% in 2001, while the "failing" rate dropped to 10% in 2000, only to rise again to 28% in 2001.

- At the Paxton Center School, although scores for eighth graders generally improved, with generally higher percentages in the "advanced" or "proficient" categories and lower percentages "failing" in both English and math over four years, scores for fourth graders were much more erratic. In English, with few students "failing," the percentage of fourth graders scoring at "advanced" or "proficient" levels declined from 48% in 1998 to 39% in 1999 and 35% in 2000 before rising to 68% in 2001. In math, while the "failing" rate dropped and remained low, the percentage of students scoring at "advanced" or "proficient" levels, rose once, from 54% in 1998 to 81% in 1999, then dropped to 56% in 2000 and 43% in 2001.

- At Boston's Hernandez School, higher percentages of fourth graders scored at "advanced" and "proficient" levels in reading and math, and the percentage "failing" math has declined steadily over four years. However, the percentage "failing" English remains inconsistent - 40% in 1998, 25% in 1999, 23% in 2000, and 31% in 2001. Eighth grade scores have also been erratic. In English, the percentage at "advanced" and "proficient" levels declined in 1998, 1999, and 2000 from 36% to 30% to 15%, then moved up to 48% in 2001. At the same time, the percentage of students "failing" English rose from 12% in 1998 to 13% in 1999 and 30% in 2000, but dropped back to 12% in 2001. In math, the percentage "failing" has declined steadily from 88% in 1998 to 36% in 2001, but the percentage scoring "advanced/proficient" has fluctuated from 0% in 1998 to 4% in 1999, 0% in 2000, and 20% in 2001.

- Although the percentage of students from Boston's Mason School who score at "advanced" or "proficient" levels has increased steadily in English and somewhat steadily in math over four years, "failing" scores have been more erratic. In English, 30% of the schools students "failed" in 1998, down to 4% in 1999, up to 15% in 2000, and down again to 3% in 2001. In math, 44% "failed" in 1998, dropping to 11% in 1999, rising to 33% in 2000, and declining again to 0% in 2001.

- After posting a jump in the percentage of students scoring "advanced" or "proficient" in English from 14% in 1998 to 37% in 1999, Worcester's Canterbury School "advanced/proficient" English rates have fluctuated dramatically, dropping to 20% in 2000, then spiking to 77% in 2001.

Four-year patterns of MCAS scores in award-winning schools show that score gains from one year to the next do not predict sustained high scores. So why do annual score gains so often fall short as indicators of school improvement? While policy makers choose to equate MCAS gains with better quality schooling, factors unrelated to authentic student achievement, including small numbers of students tested, changes in the composition of a school's testing pool, and extensive test preparation, can all push scores artificially higher regardless of school quality.

Test score gains and small numbers tested

Labeling schools as "good" or "bad" on the basis of test score gains can be especially misleading when the schools cited are testing a small number of students. In their examination of test score patterns in such schools in North Carolina and Texas, researchers Thomas Kane and Douglas Staiger (2001, March 2001; forthcoming 2002; see also Kane, Staiger, & Geppert, 2001) found that the chance occurrence of even a few "stars" or "class clowns" in the test-taking pool could skew scores dramatically from one year to the next. Moreover, spikes in scores are more dramatic when small numbers are tested.

The small number of students tested in many award-winning Massachusetts schools points to a significant flaw in using test score gains to describe school quality. Test scores may swing widely in schools of all sizes, but in general, variation in scores from year to year is much greater in smaller schools. In a recent analysis of four years of MCAS scores, Haney (forthcoming 2002) found that in Massachusetts elementary schools testing up to 100 students, math scores could vary from 15 to 20 points from year to year. In contrast, in schools testing more than 150 students, score changes from one year to the next were generally less than five points. These findings mirror those of Kane and Staiger (forthcoming 2002) who found "considerable volatility" of test scores in North Carolina schools testing 68 students or less.

In the majority of schools receiving any of the state's awards, the numbers tested are simply too small to draw conclusions that MCAS score gains signify authentic school improvement. Over time, a number of schools that have won recognition for "exceeding expectations" on MCAS one year may find themselves classified as "failing to meet expectations" the next, not because their school quality has declined but because of score fluctuations that occur naturally in schools testing limited numbers of students.

Edgerly Award-winning schools

In 11 of the 20 schools receiving Edgerly Awards in 1999, 2000, and 2001, so few students were tested - 68 or less, as suggested by the analysis by Kane and Staiger (forthcoming 2002) - that drawing conclusions about school quality is meaningless at best, irresponsible at worst. Specifically:

- Three of the five schools receiving Edgerly Awards in 1999 test fewer than 68 students on average. Over four years, the average number of fourth graders tested was 46 at Springfield's Kensington School (48 in 1998, 42 in 1999, 43 in 2000, and 52 in 2001), 54 at Boston's Roosevelt School (55 in 1998, 51 in 1999, 54 in 2000, and 55 in 2001), and 58 at Danvers's Riverside school (53 in 1998, 66 in 1999, 61 in 2000, and 53 in 2001).
- Three of the five schools receiving Edgerly Awards in 2000 test similarly small numbers of students. At the Eames Way Elementary School in Marshfield, the average number of students tested over four years was 49 (46 in 1998, 51 in 1999, 45 in 2000, and 54 in 2001). At the Lowell Middlesex Charter School, the average number tested was 39 (26 in 1999, 60 in 1999, 53 in 2000, and 16 in 2001). At the Seven Hills Charter School's, the average number of eighth graders tested was 63 (53 in 1998, 80 in 1999, 66 in 2000, and 53 in 2001).
- Even in 2001, when eight of the 10 schools receiving Edgerly Awards for score improvements from 2000 to 2001 were either high schools or vocational schools, five of the 10 schools recognized test fewer than 68 students on average. Specifically, an average of 41 are tested at Tantasqua Regional Vocational School; 45 at Worcester's Thorndyke Road Elementary School; 50 at No.Brookfield High School; 57 at Medford's Vocational-Technical program; and 60 at the Thomas Nash Elementary School in Weymouth.

MassInsight Vanguard Schools

In eight of the 12 MassInsight Corporation's Vanguard Schools, the average number of test-takers also stands at 68 or less. Again, because sharper score rises are likely when small numbers are tested, these awards may bestow the label on schools as "good" when, in fact, the schools have made few appreciable improvements in authentic student learning. Specifically:

- Sunderland Elementary School tested an average of 34 fourth-graders over four years (36 in 1998, 20 1999, 41 in 2000, and 37 in 2001).
- In the two Everett elementary schools selected as Vanguard schools, the maximum number of fourth-graders ever tested was 52. At the Albert Lewis School, the average number tested was 31 over four years (26 in 1998, 35 in 1999, 25 in 2000, and 39 in 2001). At the Devens

School, the average number tested was 48 (50 in 1998, 42 in 1999, 46 in 2000, and 52 in 2001).

- All three Woburn elementary schools selected as Vanguard schools test an average of 68 fourth graders or less. At the Altavesta School, the average number of fourth-graders tested is 33 (34 in 1998, 34 in 1999, 32 in 2000, and 36 in 2001). At the Goodyear School, the average number tested is 42 (41 in 1998, 33 in 1999, 53 in 2000, and 42 in 2001). The average number tested at the third school, the Reeves School is 68 (62 in 1998, 76 in 1999, 73 in 2000, and 61 in 2001).
- Arlington's Thompson Elementary School has tested an average of 52 students over four years (53 in 1998, 46 in 1999, 63 in 2000, and 45 in 2001).
- At Lowell Middlesex Academy Charter High School, one of only three high schools selected as a Vanguard School, the average number of students tested over four years was 39. Only 26 tenth graders were tested in 1998, 60 in 1999, 53 in 2000, and 16 in 2001.

Commonwealth Compass Schools

In the majority of schools named as Compass Schools by the Massachusetts Department of Education in 2001, the numbers of students tested are again too small to draw conclusions about either school quality or the suitability of their practices for replication. Of the 14 Compass Schools named by the state, nine, all elementary schools, tested fewer than 68 students, with several testing well below that number.

- Boston's Samuel Mason School (named as a Compass School although not listed on the Department's list of schools invited to apply for Compass school status) tested a four-year average of 27 (27 in 1998, 27 in 1999, 26 in 2000, and 29 in 2001).
- Westfield's Moseley Elementary School tested a four-year average of 39 fourth graders (32 tested in 1998, 47 in 1999, 30 in 2000, and 45 in 2001).
- Boston's Hernandez School tested a four-year average of 44 fourth graders (45 in 1998, 44 in 1999, 44 in 2000, and 41 in 2001) and an even lower average of 23 eighth graders (25 in 1998, 23 in 1999, 20 in 2000, and 25 in 2001).
- Springfield's Kensington Elementary, also an Edgerly Award winner, tested a four-year average of 46 fourth graders (48 in 1998, 42 in 1999, 43 in 2000, and 52 in 2001).
- Worcester's Canterbury Street School tested a four-year average of 52 fourth graders (51 tested in 1998, 51 in 1999, 55 in 2000, and 30 in 2001).
- Orleans Elementary School, tested a four-year average of 56 fourth graders (65 in 1998, 59 in 1999, 53 in 2000, and 47 in 2001).

- Danvers's Riverside School, also an Edgerly Award winner, tested a four-year average of 58 fourth graders (53 in 1998, 66 in 1999, 61 in 2000, and 53 in 2001).
- Salem's Saltonstall Elementary School tested a four-year average of 59 fourth graders (56 in 1998, 52 in 1999, 56 in 2000, and 71 in 2001).
- Paxton Center Elementary School tested a four-year average of 63 fourth graders (56 in 1998, 68 in 1999, 68 in 2000, and 60 in 2001) and 50 eighth graders (41 in 1998, 48 in 1999, 53 in 2000, and 60 in 2001).

Elementary schools and charter schools are most likely to test small numbers of students, making the use of test score gains to cite "high quality" schools most troubling in these categories.

- Of the 179 elementary grades schools on the state's list of "exemplary" schools, 116, or two out of three, test fewer than 68 fourth graders each year.
- All four of the secondary charter schools cited as exemplary - two in eighth grade, two in tenth grade - test fewer than 68 students, with the average annual number tested ranging from only 21 at South Shore Charter High School to 39 at Lowell Middlesex Academy Charter School for the four years of testing.

Middle and high schools typically test larger numbers of students. However, even in these schools, at least half the variation in scores from one year to the next typically reflects what researchers call "noise" attributed to factors unrelated to authentic student achievement (Kane & Staiger, forthcoming 2002, Table 2). Of the 38 district middle grades schools invited to apply for Compass School status, six test fewer than 68 on average each year. Among the 20 district high schools in this category, Provincetown tests an average of 29 tenth-graders annually, and nine typically test 100 students or less.

The hazards of confusing MCAS gains with good practice

Award-winning schools may be "good schools," but MCAS scores hardly provide conclusive evidence for such claims. Given wide score swings that occur as a matter of course in schools testing low numbers of students, similar schools where scores decline may be equally "exemplary." And given that so many schools test small numbers of students, educators from either group of schools could eventually find themselves defending score lapses that occur for no

reason other than chance. If evidence from other states indicates what could happen in Massachusetts, pressure to produce MCAS gains could also result in some schools abandoning promising, but long-term, reform efforts in favor of activities that produce a "quick fix" (Frommeyer, 1999).

Ultimately, natural score volatility will "wreak havoc" with accountability systems as educators are rewarded or punished for score fluctuations that occur due to conditions beyond their control (Kane & Staiger, March 2001; forthcoming 2002). Assuming that schools ranked as "exemplary" can necessarily guide others toward better practice adds to the problem. As Kane and Staiger (March 2001: 2; forthcoming 2002: 2) write:

> To the extent such rankings are used to identify best practice in education, virtually every educational philosophy is likely to be endorsed eventually, simply adding to the confusion over the merits of different strategies of school reform. For example, when the 1998-99 MCAS test scores were released in Massachusetts in November of 1999, the Provincetown district showed the greatest improvement over the previous year. The *Boston Globe* published an extensive story describing the various ways in which Provincetown had changed educational strategies between 1998 and 1999, interviewing the high school principal and teachers.

Since school scores vary more dramatically than statewide scores, scores in individual schools, especially small ones, do rise more dramatically than scores statewide or in larger schools, creating the impression that they are accelerating their students' achievement. As a result, policy makers, supporters of test-based "accountability," and the media may mistakenly endorse whatever practices evident in such schools. As Kane and Staiger explain, "If school-level test scores are the gauge, the *Boston Globe* and similar newspapers around the country will eventually write similar stories praising virtually every variant of educational practice" (March 2001: 3; forthcoming 2002: 3).

Since score gains look most dramatic when small numbers are tested, awards programs based on MCAS score gains are weighted toward small schools. In future award cycles, additional Massachusetts schools testing small numbers of students may be cited as "exemplary" when, in fact, score gains are based on chance or statistical patterns associated with a small testing pool, not genuine improvement. As David Grissmer of the RAND Corporation reflects on school awards policies, "The question is, are we picking out lucky schools or good schools, and unlucky schools or bad schools? The answer is, we're picking out lucky and unlucky schools" (Olson, 2001: 9).

Rising test scores and the changing composition of students tested

> *Anytime you have groups of different kids taking the test each year, you're going to have different results. The scores are going to change each year because they're different kids. If the curriculum stays the same, and the teachers stay the same, but the results change, it's the students.* - Stuart Peskin, principal of Bennett-Hemenway School in Natick, a school that exceeded Department of Education goals for MCAS score gains, quoted in Miller, 2001).

In the haste to claim that high stakes testing produces better schools, policy makers often overlook the reality that comparing "apples" to "oranges" - stacking the scores of students tested in one year against those of students tested the next - misleads the public about the meaning of score gains. Indeed, this misuse of test scores has undermined credibility of school accountability programs in other states. As Ken Jones of the University of Alaska and Betty Lou Whitford of Columbia University (1997: 278) explain:

> Significant controversy arises from the fact that, in determining whether or not a school is making progress, different groups of students are tested each year. That means, for example that one group of fourth graders is being compared with a different group of fourth graders.

In fact, score gains in Massachusetts award-winning schools may result from the simple fact that a particular cohort of students contains stronger students than cohorts from prior years.

Scores can also rise because of demographic changes in the larger community, or when weaker students do not participate in testing, either because they are retained in the grade prior to testing and are not tested with their cohort, or because they have left school altogether.

Ninth grade retention

From one year to the next, a decline in low-scoring students in the grade tested may occur as a result of chance circumstances. But particular school practices and policies may also change the characteristics of student test-takers and help boost test scores in particular schools. When schools retain more students in grade, especially in the years prior to testing, or when more students are absent from the roster of test-takers, score gains may owe more to the loss of low-scoring students from the population tested than to improvements in teaching and learning (Allington, 2000; Allington & McGill-Franzen, 1992; Darling-Hammond, 1997; Elmore, 1997; Haney, 2000; Haney, 2001; Jones, 2001; McGill-Franzen & Allington, 1993).

Since MCAS scores were declared the basis for assessing schools, statewide data have registered an increase in grade retention in the ninth grade, rising from 6.8 percent of ninth graders retained in 1997-98, to 7.4 percent in 1998-99, and to 8.1 percent in 1999-2000 (McElhenny, 2001a). At best, ninth grade retention reduces the number of low-scoring students taking MCAS the following year. At worst, it discourages vulnerable and overage-for-grade students from persisting in school through tenth grade.

When more students are held back in grade nine, MCAS scores can get a boost in grade ten. Of the seven high schools and vocational schools (excluding the Medford Vo-Tech program where numbers were too small to be meaningful) receiving Edgerly School Awards for 2001 MCAS scores, increases in ninth grade retention rates from 1999 to 2000 likely helped push down "failing" rates on MCAS in 2001. For example:

- Boston's Charlestown High School's ninth grade retention rate jumped from 6.4 in 1999 to 11.5 in 2000. The school's percentage of tenth graders "failing" MCAS dropped from 84% in English and 81% in math in 2000, to 41% in English and 39% in math in 2001.
- Gateway Regional High School's ninth grade retention rate increased from 13.3 in 1999 to 16.8 in 2000. The school's percentage of tenth graders "failing" MCAS dropped from 46% in English and 68% in math in 2000 to 16% in English and 26% in math in 2001.

In three of the 20 district high schools invited to apply for Compass School status on the basis of MCAS score gains from 1998 to 2000, "failing" rates dropped over three years while high ninth grade retention rates in 1998 and 1999 removed weaker students from those tested in tenth grade in 1999 and 2000. Specifically:

- Ayer High School retained 13.8% of its ninth grade in 1998, 19.8% in 1999. From 1998 to 2000, the percentage of tenth graders scoring "failing" dropped from 19% to 12% in English and from 47% to 26% in math.
- Southbridge High School retained 18.6% of its ninth grade in 1998; 19.8% in 1999. From 1998 to 2000, the percentage of tenth graders scoring "failing" dropped from 34% to 25% in English and from 54% to 38% in math.
- Ralph C. Maher High School retained 9.7% of its ninth graders in 1998, 13.4% in 1999. From 1998 to 2000, the percentage of tenth graders scoring "failing" dropped from 37% to 26% in English and from 57% to 41% in math.

Higher rates of retention in ninth grade are a source of concern not only because they artificially boost MCAS scores but also because grade retention undermines student achievement while contributing to dropping out (Heubert & Hauser, 1999; Wehlage & Rutter, 1986; Smith & Shepard, 1989). Holding more students back in the grades prior to testing may improve school scores in the short run, but over time, individual student achievement will not improve, and a larger portion of the state's dropouts, many of whom are already overage for grade, will leave school with less than a tenth grade education.

Dropping out and MCAS higher scores

As tenth grade MCAS scores have improved statewide in Massachusetts, the Massachusetts dropout picture has also shifted. Although the state's annual high school dropout rate has hovered between 3.4 and 3.6 in the early years of MCAS testing (Massachusetts Department of Education, 2001b), an analysis of state data shows that more Massachusetts dropouts are leaving school in the ninth and tenth grades, even before taking MCAS. In 1997-98,

49.9% of the state's 8,582 dropouts were ninth or tenth graders; by 1999-00, 54.3% of the state's 9,199 dropouts came from these grades.

In nearly one third - 10 out of 33 - of the high schools and vocational schools (excluding two charter schools) that have won awards and recognition for MCAS score gains, dropout rates are higher now than in 1997, the year before MCAS testing began.

- Of the 11 high schools or vocational schools receiving Edgerly Awards, four posted higher dropout rates in 2000 than in 1997, the year before MCAS testing began. These include Gateway Regional High School, Swampscott High School, Medford Vocational-Technical High School, and Tantasqua Regional Vocational High School. For example, the annual dropout rate at Gateway Regional High School has increased steadily from 3.3 in 1997, to 4.6 in 1998, 4.8 in 1999, and 6.3 in 2000, nearly twice the state rate of 3.4.
- Of the two high schools receiving MassInsight Vanguard Awards, both posted higher annual dropout rates in 2000 than in 1997, before MCAS testing began. Hudson High School's annual dropout rate was 1.5 in 1997, up to 2.6 in 2000. Nauset Regional High School's annual dropout rate was 1.4 in 1997, up to 2.7 in 2000.
- Of the 20 high schools invited to apply for Compass School status on the basis of MCAS score gains, five posted higher dropout rates in 2000 than in 1997. Although numbers at these largely white schools remain small, Ayer High School, Boston Latin School, Hudson High School, Provincetown High School, and Swampscott High School posted higher annual dropout rates in 2000 than in 1997.

Dropout rates among award-winning schools underscore the reality that MCAS score gains alone are poor means of identifying "good" schools. In a state committed to improving learning for all students, schools with high rates of retention and rising annual dropout rates should not be considered "exemplary" simply because MCAS scores rise. Indicators of school holding power and inclusion must be considered as well.

Disappearing tenth graders

Schools' MCAS scores can also rise if weaker students "disappear" between October and May of their tenth grade year. Although this loss of tenth graders could reflect the movement of families out of the state, students enrolled in October of a school year may "go missing" from MCAS testing in May for a number of other more troubling reasons. Some may officially drop

out of school. Others may transfer into a private or parochial school in their community. Regardless of the reason, an increase in the percentage of tenth graders who go "missing" between October and May of the school year can change the population of tenth graders tested from one year to the next and boost a school' s MCAS scores.

In October 2000, 68,577 students were enrolled in tenth grade in Massachusetts; in May 2001, approximately 62,000 tenth graders took the MCAS. The loss of some 9.6% of the state's tenth graders between October and May is about the same as that of the two previous school years. However, in some award-winning schools, the percentage of "missing" tenth graders has not remained stable, but has moved steadily higher.

- In two of the Vanguard Award-winning high schools, the percentage of students "missing" in tenth grade increased steadily from 1998 to 2000, the year the schools were named Vanguard Schools. At Hudson High School, 18.2% of tenth grades enrolled in October 1997 (29 out of 159) did not take the MCAS in May 1998; 24.1% of tenth graders enrolled in October 1998 (39 out of 162) did not take the MCAS in May 1999, and 29.9% of tenth graders enrolled in October 1999 (47 out of 157) did not take the MCAS in May 2000. At Nauset Regional High School, 1.6% of tenth graders enrolled in October 1997 (4 out of 256) did not take the MCAS in May 1998; 6.5% of tenth graders enrolled in October 1998 (16 out of 247) did not take the MCAS in May 1999; and 8.5% enrolled in October 1999 (21 out of 248) did not take the MCAS in May 2000.

- Of the 20 district high schools invited to apply for Compass School status for test score improvements posted for 1998, 1999, and 2000 (and where the numbers tested were higher than 30), six others (Carver, Clinton, Manchester, Ware, Oakmont, and Ralph C. Mahar) also had higher rates of October-to-May loss in the 1999-2000 school year than in the 1997-98 school year. Among these schools, the smallest loss was at Oakmont Regional High School, where 5.3% of the tenth graders enrolled in October 1999 (10 out of 190) were not tested in May 2000. The largest loss was at Clinton High School, where 19.0% of the students enrolled in tenth grade in October 1999 (24 out of 137) were not tested in May 2000.

State data sources propose no explanation for the increases in missing tenth graders in particular communities, but these increases in a number of award-winning schools highlight how test score gains, on their own, reveal very little about school quality. In fact, awards programs may discourage schools from holding on to students whose test-score prospects threaten schools' rankings. Some schools may enroll students as tenth graders in October, but delay their

participation in MCAS testing for a year by reassigning them to ninth grade homerooms. Others may limit the personalized attention or diversified instruction that vulnerable students may need to prevent them from dropping out. Some may simply have increasing numbers of English-as-a-second-language students who are not tested during their first years as newcomers to a school. In some districts, the threat of the denial of a high school diploma may encourage parents with means to remove students from the testing pool and enroll them in a private, parochial, or home school for the final years of high school.

Clearly school awards programs fail to account for the multiple ways in which test scores can rise when weaker students are not included in the testing pool, whether on a permanent or temporary basis. Indeed, programs that recognize schools primarily for MCAS gains and distribute monetary awards to particular schools named "exemplary" provide little incentive for schools or the public to explore all possible reasons for score gains, including the routines and practices that might change the population of students tested. Under pressure to meet or exceed score expectations, schools may find ways to *look good* without necessarily developing greater capacity to engage all students in authentic learning. But when ninth grade retention rates rise and the percentage of students leaving school increases, MCAS score gains are cause for worry, not celebration.

Test preparation in school, out of school, on-line: Valuing scores more than learning

> *I can't make you smarter. All I can do is help you take the test better, so that's what I'm going to do.* - Teacher Joseph Saia, to students in an after-school MCAS preparation session, quoted in Greenberger & Vaishnav, 2001: B7.

Reports from across Massachusetts suggest that schools are devoting increasing amounts of instructional time to test preparation, both during the regular school day and in Saturday, after-school, and summer school classrooms (see, for example, Astell, 2001; Berkley, 2001;

Doherty, 2001; Cameron, 2001; Connolly, 2001; DeForge, 2001; Gonter, 2000, 2001; Greenberger & Vaishnav, 2001; Gutstein, 2001; Huang, 2001; Massey, 2001a, 2001b; Myers, 2001a; Nichols, 2001; Wicka, 2000). As teachers focus on coaching students in test-taking skills, open-response items that are "carbon copies" of MCAS questions are becoming a routine part of the curriculum, replacing project work in student portfolios in favor of mock time trials on MCAS questions. Some schools have hired extra teachers specifically for in-school MCAS instruction. While teachers in affluent districts turn their classes into MCAS preparation periods a week before the tests, those in lower-income districts set up year-round MCAS "review" classes for students deemed at risk of failure, a label that applies to more than half the students in a given grade in some schools. Vacation-time test preparation classes walk students through practice problems and alert students to test instructions and formatting issues.

Test companies do not view MCAS as as "coachable" as the SATs, but they argue that MCAS preparation programs can equip students with test-taking strategies (Greenberger, 2001). To this end, some districts have redirected resources toward the purchase of test-prep materials; others have hired private companies to make on-line test-prep software available to all students and for use in tutoring students at risk of failing MCAS (Massey, 2000; Wilson, 2001; Myers, 2001b). To help districts identify such resources, the Massachusetts Department of Education maintains information for schools on commercially-prepared programs and works with some commercial vendors to reduce the costs of products and services for public schools (Massachusetts Department of Education, 2001).

In the context of high stakes testing that focuses attention on test score gains, scores may indeed improve as teachers and students become more familiar with the format and content of high stakes tests, and as teachers devote an increasing amount of classroom time to drilling

students on test-taking skills (Cuban, 2001; Hoff, 2000; Kohn, 1999; Koretz, 1988; Madaus & Clarke, 2001; McNeil & Valenzuela, 2001; Smith & Rottenberg, 1991). However, although such test preparation may produce higher scores in the short term, gains posted as a result of test preparation for one test rarely generalize to performance on other tests (Koretz, Linn, Dunbar, & Shepard, 1991). Moreover, when schools set annual score gains as a primary goal, content in areas other than those defined by the tests may be sacrificed (Kohn, 2001). Tailoring class work to fit the content of MCAS test questions, schools have made changes as simple as replacing the study of Shakespeare's *MacBeth* with *A Midsummer Night's Dream* (Hoboth, 2000). But to make time to prepare students for MCAS, schools have also dropped or de-emphasized courses in science, American government, black history, and physical education, that some would argue are essential to student growth and development as healthy citizens (Hagan, 2000; Hand, 2001; Hayward, 2001b; Rene, 2001). At Lowell High School, lunch periods now begins at 9:25 to accommodate a new schedule that squeezes a new MCAS prep seminar into the school day (Lipman, 2001; Scarlett, 2001).

Under pressure to produce higher scores, educators may work harder to achieve measurable, targeted goals and the rewards that accompany them. However, as teachers turn to more controlling instructional strategies and focus their expectations for student learning on ensuring that students "get the right answer" on state tests, students' motivation and development as independent learners is in jeopardy (Paris & Urdan, 2000). Kennon Sheldon and Bruce Biddle (1998: 176) explain the paradox: "Although maximal student growth may be the goal, if student attention is focused on tests that measure that growth, or on sanctions that reward or punish it, that growth will not be maximized." University of Michigan researcher Scott Paris (2000) adds:

> When test scores and grades define educational success, students value the outcomes more than the knowledge or processes of learning. This occurs for the high scorers as well as the low scorers and demeans education for all students. Indeed, it is a serious threat to life-long learning and the recreation of discovery that is so important in a world that demands continual effort to keep pace with growing technology, international news, and community opportunities.

School awards and recognition programs that cite schools as "exemplary" are intended to highlight "best practices" that can be replicated from school to school. But if test preparation is the engine behind test score gains, and if schools devote increasing amounts of time to producing better score results, authentic "best practice" may be hard to identify in such schools. And although the teaching of test-taking strategies may boost scores in the short term, gains eventually level off, even in authentically good schools. As Harvard professor Daniel Koretz notes, "The notion that there will be continuous improvement is a little optimistic at best. You can teach them more, and you can teach them faster, but at some point, you're going to top out" (Hoff, 2000:19).

Accountability that strengthens schooling for all

Massachusetts accountability and school recognition policies fail to identify in any holistic or authentic way which Massachusetts schools are "more exemplary" than others and, at the same time, have harmful consequences. First, these policies narrow the definition of "exemplary schooling" by ignoring the multiple dimensions of what constitutes good schools. Americans have traditionally wanted schools to develop children's intellect, but they also expect schools that meet goals for social, vocational, and personal development (Goodlad, 1984). As researchers have long emphasized, test scores do not assess schools' capacity for generating students' curiosity and disposition to ask probing questions, engaging student motivation, developing skills in working as a team, or setting norms for positive interaction between teachers and students (Madaus, 1983).

The reliance on MCAS score gains to point to "exemplary" schools has a chilling effect on authentic school reform for another reason: Allowing MCAS score gains, rather than indicators selected by the local community from a broader menu, to dominate accountability practice means that schools are not challenged to develop a sustained capacity for improving school conditions and their own "best practice" in ways that lead to performance-based achievement and strengthen schools as inclusive learning communities. Ultimately, parents, educators, and decision-makers who seek deeper understanding about the work of schools need information about the quality of work students do, students' access to resources and opportunities to learn, and conditions that foster a press for achievement and professional practice in their schools (Oakes, 1989). The MCAS-based awards programs provide the public with little information about these aspects of school improvement, limiting the tools educators and parents need to effect authentic and holistic schooling that benefits all students.

If top-down test-based accountability models do not provide reliable signposts for improvement, what shape should an alternative accountability policy take? A redefined approach to school accountability would derive from six basic principles proposed by the Massachusetts Coalition for Authentic Reform in Education (Massachusetts Coalition for Authentic Reform in Education, ND):

- No single assessment tool can adequately assess schools or student learning. Test scores are only one source of information for improving student achievement; student work, not test scores, should be used to gauge the quality of student learning and the assignments that students receive.
- Accountability should go beyond a "test scores only" approach to require schools to "account for" the practices they employ during the school day that strengthen teaching, engage all students in learning, and ensure students will produce work that reflects high standards of quality. Data on dropout rates, grade retention, attendance, and suspension are also essential for painting a picture of schools' sense of accountability for all students.

- Standards for quality student work must be set at the local level through a partnership between the state and local communities.
- Professional accountability depends on the opportunities educators have to define school goals and work in collaboration with other professionals, parents, and community members toward achieving those goals.
- Accountability requires the state to monitor, protect, and expand access of all students to high-quality equal learning opportunities and resources. The state is responsible for providing technical assistance to schools to correct practices that undermine equal access to learning, including such practices as tracking, grade retention, and punitive attendance and suspension policies.
- Reviewers from outside the school and district play a key role in ensuring credibility of a school accountability system.

CARE's proposal for school accountability makes use of test scores, but also assumes whole communities will focus on student work, not test scores, as the touchstone for discussing standards-based practice. The proposal focuses on student work through review of student portfolios, projects, and presentations. It emphasizes the strengthening of professional practice so as to improve the decisions teachers make about curriculum and instruction in their own schools; help school communities dissect their own problems and learn from mistakes; and establish an ethos wherein teachers take responsibility for the progress of all students (Darling-Hammond, 1997; Dorn, 1998; Haney & Raczek, 1994; Sirotnik & Kimball, 1999). It also expects that the state will act to ensure that all schools have the resources necessary to ensure that all students have equal access to high-quality opportunities to learn (Elmore, 1997).

CARE's proposal for an alternative approach to school accountability calls for a multifaceted system designed to promote high standards for learning without high stakes testing. It assumes that local schools know their students best, and therefore, that the state's role is not to make decisions about individuals. Rather, the state's responsibility is to ensure that schools are educating all children well and to provide sufficient resources and assistance to enable schools to do so.

CARE's proposal avoids the pitfall of relying on a single assessment to meet a range of goals by integrating multiple assessments designed for different purposes into a coherent whole. Limited statewide standardized testing in reading and math would monitor student achievement at the state and district level. Locally developed performance-based assessments tied to state education goals would provide information on individual student learning. School quality reviews would provide school-level information about teaching and learning that schools and districts can use for school improvement. School reports to the community would provide information to parents and community members are informed about district, school, and student performance in relation to standards for achievement, resource allocation, equity, and holding power.

Limited standardized testing in literacy and numeracy

Limited standardized testing in literacy and numeracy is one tool in an accountability program oriented to school improvement. Such testing is a useful tool for monitoring student performance in reading and math statewide, by district, and by race. In the past, Massachusetts gathered such information through the Massachusetts Educational Assessment Program (MEAP) administered in selected grades. Similar to the National Assessment of Educational Progress (NAEP), MEAP also gathered information about students' opportunities to learn and perceptions of their schooling experiences. Testing for monitoring state and district performance should be administered in a way that imposes the least burden possible on districts and intrudes to a minimal extent on teaching and learning.

Local assessments based on the Massachusetts Common Core of Learning and developed in the districts

Many Massachusetts districts already administer national norm-referenced tests in at least some grades. CARE's proposal calls for each district, working with professionals at each school, to supplement such testing with local assessments designed to help teachers improve instruction

and assess the performance of individual students by focusing on student work, including projects, portfolio reviews, and presentations. Researchers Monty Neill and Keith Gayler (2001) report that by strengthening teachers' capacity to use such assessments as part of classroom life, locally developed assessments hold strong promise for strengthening the achievement of all students, including that of traditionally low-scoring students.

Local assessments would require students to demonstrate skills and understanding of content as defined within the broad parameters of the Massachusetts Common Core of Learning and streamlined state curriculum frameworks. Local schools councils, along with district and state leaders, will review and approve school assessment and accountability plans, including rubrics and exemplars of high quality work, a description of how students' work quality will be reported to parents, and criteria for graduation and promotion. Teachers will be responsible for making graduation decisions based on multiple criteria.

School quality reviews

School quality reviews (SQRs) complement data provided from student assessments by providing in-depth information about teaching and learning in every school. As the third component of CARE's accountability proposal, SQRs represent a key strategy for moving beyond assessing "outcomes" to examining the daily learning experiences students have during the school day, teaching practices, and the quality of student work in relation to expected standards of quality. SQRs are also key to developing schools' capacity to review their own practice and to work in partnership with professionals from outside the school to learn from their strengths and weaknesses.

A wealth of experience is available to guide the state in facilitating schools' engagement in school quality reviews geared to develop the professional capacity of educators across the state.

Schools seeking accreditation from the New England Association of Schools and Colleges already undergo an intensive evaluation every 10 years, involving an in-depth self-study, a four-day visit by a team of 12-14 educators that assesses the school in terms of its own goals and standards. Massachusetts schools that belong to school reform networks like the Coalition of Essential Schools likewise engage in rigorous school reviews in partnership with professionals from other schools in the Coalition. As organizer and facilitator of school quality reviews, the state Department of Education can draw from these resources as well as the long-standing experience in England, Ireland, and Scotland where school inspectorates represent the primary tool for school standards-setting and accountability or on more recent experience in New York State (Ancess, 1996; Wilson, 1996).

School quality reviews in Rhode Island provide the closest example of the way in which a state department of education can make professionally-grounded school reviews the cornerstone of an accountability approach that used school assessments to help schools develop their capacity for ongoing improvement. By law, the Rhode Island Department of Education's School Accountability for Learning and Teaching (SALT) office, working with the state's Field Services Office, Office of Progressive Support and Intervention, and educational networks and collaboratives, is responsible for developing and implementing systems that support the continuous improvement of schools. In practice, this mandate translates into a school quality review process.

SALT initiates this process by requiring schools to form school improvement teams, then working with them to engage in a self-study based on an analysis of student assessment data and results of parent, teacher, and student surveys. Based on the self-study, schools then develop an

improvement plan for improving student performance and present the plan at an annual school report night.

In addition, once every five years, each school must host a SALT visit that follows procedures developed in visits to 123 schools since 1997-98. Teachers-on-leave to the Rhode Island Department of Education serve as trained SALT fellows and chair visiting teams. Other team members include principals, diverse subject area teachers, and librarians, staff from the Rhode Island Department of Education, and parents from outside the district. All are trained to gather information and make their observations through a professional lens.

Similar to an accreditation visit, the SALT visit lasts over several days to a week and focuses on teaching, learning, and the school climate and operations. The team conducts extensive observations of classrooms and teacher planning time, reviews documents like the school's strategic plan, meets with the school's improvement team, students, parents, teachers, and school administrators, draws on data from the SALT survey, to answer such questions as: Does the school's plan have adequate focus to accomplish its mission and goals? How effective is the school's communication with families? Are the school's instructional programs sufficient to equip student to meet the school's performance targets?

Following the visit, the SALT team makes its report to the school improvement team, including final recommendations. These may acknowledge special problems the school faces - whether high transience or inadequate facilities - but they also emphasize that the school must solve identified problems related to curriculum, instruction, expectations for students, or quality of support services. Finally, the entire report is certified by an outside "endorser," a professional who has observed at least part of the visit and is knowledgeable about school quality reviews, based on specific criteria and procedures certifies that the report was properly produced and

conducted in the expected manner. (For more detailed information, see Rhode Island Department of Elementary and Secondary Education, ND: http://www.ridoe.net/schoolimprove/salt/faqs.htm. For sample reports, see postings at the Rhode Island Department of Education web site: http://www.ridoe.net/schoolimprove/salt/visit/vismenu.html).

CARE's proposal calls for a each school in Massachusetts to engage in a SQR process similar to that of Rhode Island every four to five years. Each school would be required to begin the process with a detailed self-study. An expert team would conduct an extended visit to the school to interview students, educators, and parents, observe classes and teacher meetings, and review examples of student work and school policies. At the end of the visit, the review team would present a face-to-face report to the school, followed by a detailed written report within a month of the visit. Teams would be organized by the Department of Education or developed in collaboration with the regional accreditation association. Reports could also trigger further in-depth technical assistance by the state directed toward school improvement.

Annual reporting by schools to their communities

Ultimately, accountability involves reporting not only on results, but on actions taken in relation to results. Because school accountability, then, requires professionals in explaining their practice to their community, CARE proposes that each school in the Commonwealth present annual reports on both school progress and practice to parents and the larger community. Formal reports would address school practice in relation to academic learning and state curriculum frameworks and would include results of test scores along with examples of student work. Reports would also include information about steps the school is taking in relation to achieving its own goals related to school holding power, opportunities and resources to learn, routines that expose all students to a "press for achievement," and teachers' own opportunities to

develop collegial practice. Reports would include outcomes by race, special education, limited English proficient, and Title 1 status, when appropriate, and be prepared in collaboration with external partners familiar with school goals and operations (see, for example, Center for Policy Analysis, University of Massachusetts, 2000). Local school councils, parents, and other community members, the district, and state would review reports. When needed, the state or district can send in teams to verify the accuracy of a school's report.

CARE anticipates that school reports to parents and the community would also include a variety of forums to highlight student achievement. Student-led parent conferences, "Culminating nights," and review panels where exiting students present their work to parents, school committee members, and others from the community are all ways in which schools expand parents' and communities' understanding of the standards schools set for the quality of student learning (Berger, 1996; Frommeyer, 1999; Wheelock, 1998).

The CARE plan opens up multiple opportunities for reviewing the quality of student work and classroom practices. As such, it provides a multidimensional view of school operations. Most important, it provides a baseline of information from which educators and community members, in partnership with the state, make decisions about school strengths, weaknesses, and steps toward continuous improvement. Given this foundation, the state will have higher quality information so that if intervention is necessary, it can fashion a school improvement plan that addresses both student achievement and practices within the school.

Conclusion

Massachusetts accountability policy, including its MCAS-based awards programs, assumes that public reporting of MCAS score gains is the key to school improvement. "It is this test, even more than the nearly $6 billion in new funds, that will be the real impetus to improve

our schools," Mark Roosevelt, former state senator and education reform legislation architect, has said (McFarlane, 2000). "The School and District Accountability System is the shining star of education reform in that it's taking schools for what they are, where they're starting off, and allowing them to show what they can do in terms of improvement," says state Department of Education spokesman Jonathan E. Palumbo (Tantraphol, 2001). Acting Massachusetts Governor, visiting Westfield's Moseley Elementary School, proclaims, "You must be doing something right if you are a Compass School" (Malay, 2001).

However, rhetoric does not always reflect reality. MCAS score gains are not the valid or reliable indicators of school improvement that policy makers imagine. Nor are they necessarily signs that schools that are "doing something right." In many Massachusetts schools listed as "exemplary," statistical patterns associated with small numbers of students tested, changes in the composition of a school's students taking the MCAS from one year to the next, and teaching to the test may artificially improve test scores without improving school quality. By using MCAS score gains to identify particular schools as models of schools improvement, public policy makers and pro-MCAS corporate leaders promote an inadequate definition of school quality, misrepresent schools cited for test score gains as more "exemplary" than others, and do a disservice to parents, teachers, and students who seek authentic school improvement and who care more about public education than public relations.

Current accountability policies in Massachusetts, including test-based awards programs, mislead the public into believing that test score gains are fair and accurate measures of school improvement. This top-down, test-based approach to school assessment and accountability should be replaced in favor of a system of authentic accountability. CARE's proposal aims to develop each school's capacity to assess and "account for" the quality of education provided to all students

through a process that balances results from external tests and reviews with locally-based assessments of student work. This proposal calls for a combination of standardized testing to monitor student skills statewide, local assessments that focus on student work, professionally-organized reviews of school quality, and a reporting process that requires educators to describe student learning outcomes and opportunities to their own community in the context of their schools' organization and practice. This approach, rather than a top-down, MCAS-driven school accountability policy is key to making "accountability" one aspect of a larger commitment to education reform that benefits all students.

References

Allington, R. L. & Mc-Gill-Franzen, A. (1992). Does High-stakes Testing Improve School Effectiveness? ERS Spectrum, 10 (2): 3-12.

Allington, R. L. (2000). How To Improve High-stakes Test Scores Without Really Improving. Issues in Education: Contributions from Educational Psychology, 6, (12): 115-124.

Ancess, J. (1996). Outside/Inside, Inside/Outside: Developing and Implementing the School Quality Review. New York: National Center for Restructuring Education, Schools, and Teaching (NCREST), Teachers College, Columbia University.

Astell, E. (2001). "One-on-one tutoring is MCAS strategy," Worcester Telegram, 12 July: http://www.telegram.com/news/page_one/tutors.html.

Bauer, S. (2000). "Should Achievement Tests Be Used to Judge School Quality?," Educational Policy Analysis Archives, 8 (46), 17 September: http://epaa.asu.edu/epaa/v8n46.html.

Berger, R. (1996). A Culture of Quality: A reflection on practice. Providence, RI: Annenberg Institute for School Reform.

Berkley, J.M. (2001). "Affluent schools make the grade," Brookline TAB, 18 January: http://www.townonline.com/brookline/news/topstories/general/989881_0_affluent_011801_9c1244d664.html.

Blair, J. (1998). "Pa. Doles Out $10 million To Reward Schools," Education Week, 18 (8), 21 October: 5.

Bolon, C. (2001). "Significance of Test-based Ratings for Metropolitan Boston Schools," Education Policy Analysis Archives, 9 (33), 12 September: http://epaa.asu.edu/epaa/v9n33.html.

Bradley, A. (1996). "Model Chicago Schools To Get Cash Awards," Education Week, XV (17), 17 January: 1, 12.

Cameron, J. (2001). "'M' camp helping with skill sharpening," Daily Hampshire Gazette, 11 July: http://www.gazettenet.com/07072001/schools/3950.htm.

Camilli, G. & Bulkley, K. (2001). Critique of 'An Evaluation of the Florida A-Plus Accountability and School Choice Program.' Education Policy Analysis Archives," 8 (46), 4 March: http://epaa.asu.edu/epaa/v8n46.html.

Caruso, D. B. (2001). "Many schools fail in MCAS improvement," Metrowest Daily News, 10 January: http://www.townonline.com/metrowest/news/state/989889_0_many_011001_453c714b42.html

Center for Policy Analysis, University of Massachusetts. (2000). Sandwich Public Schools Community Report Card. Dartmouth, MA: University of Massachusetts.

Connolly, D. (2001). "Schools hire specialists to boost MCAS scores," Brockton Enterprise, 12 January: http://www.enterprise.southofboston.com/display/inn_news/News/news03.txt.

Cuban, L. (2001). "Two decades of school reforms take us back to the 1950s," Los Angeles Times, 18 February: 21.

Darling-Hammond, L. (1997). The Right to Learn: A Blueprint for Creating Schools That Work. San Francisco: Jossey-Bass.

DeForge, J. (2001). "Test preparations to increase odds," Springfield Union-News, 6 December: http://www.masslive.com/chicopee/unionnews/index.ssf?/news/pstories/ch126sch.html.

Doherty, J. (2001). "School chief has list of changes," New Bedford Standard-Times, 3 August: http://www.s-t.com/daily/08-01/08-03-01/a04lo021.htm.

Dorn, S. (1998). The Political Legacy of School Accountability Systems. Education Policy Analysis Archives, 6 (1), 2 January: http://epaa.asu.edu/epaa/v6n1.html#weaknesses.

Elmore, R.F. (1997). "The politics of education reform," Issues in Science and Technology, 14 (1), Fall: 41-49.

Frommeyer, S. J. (May, 1999). Implementing an Integrated Portfolio In a Kentucky Middle School: The Courage to Sustain Change and Its Implications for Leadership and Accountability. Doctoral Dissertation. University of Louisville.

Gehring, J. (2001). "Mass. Conference Examines Schools in 'Vanguard,'" Education Week, 20 (39), 6 June: 3.

Gonter, N. (2000). "MCAS hour part of drill," Springfield Union-News, 20 December: http://www.masslive.com/newsindex/holyoke/index.ssf?/news/pstories/ho1219ho.html.

Gonter, N. (2001). "Summer students focus on MCAS," Springfield Union-News, 2 August: http://www.masslive.com/holyoke/unionnews/index.ssf?/news/pstories/ho82cent.html.

Goodlad, J. I. (1984). A Place Called School: Prospects for the Future. New York: McGraw-Hill.

Greenberger, S. S. & Vaishnav, A. (2001). "Mastering MCAS," Boston Sunday Globe, 18 November: B1.

Greenberger, S. S. & Vaishnav, A. (2000). "State's new school ratings raise concerns," Boston Globe, 14 December: A01.

Greenberger, S. S. (2001). "Test-prep firms eye Swift MCAS plan," Boston Globe, 31 August: B5.

Gutstein, P. (2001). "Preparing to retake MCAS," Daily Hampshire Gazette, 3 December: http://www.gazettenet.com/12032001/schools/9176.htm.

Hagan, S. (2000). "MCAS scuttles course in American government," Danvers Herald, 13 January: http://www.townonline.com/north/danvers/news/0-9886_0_mcas_011300_3018c94cb6.html.

Hand, J. (2001). "MCAS hurts other classes," Attleboro Sun Chronicle, 27 February: http://www.thesunchronicle.com/display/inn_city/city2.txt.

Haney, W. (Forthcoming 2002). Lake Woebeguaranteed: Misuse of test scores in Massachusetts. Education Policy Analysis Archives.

Haney. W. (2000). The Myth of the Texas Miracle in Education, Education Policy Analysis Archives, 8 (41), 19 August: http://epaa.asu.edu/epaa/v8n41.

Haney, W. M. (2001). Letter to Senator Edward M. Kennedy, 30 July.

Haney, W. & Raczek, A. (1994). Surmounting outcomes accountability in education. Report prepared for the U.S. Congress Office of Technology Assessment.

Hartzel, P. (2001). "Southeastern Regional High plans MCAS test prep classes," Sharon Advocate, 26 January: http://www.townonline.com/neponset/sharon/news/03752813.htm.

Hauser, R. (2001). Should We End Social Promotions?: Truth and Consequences. In Orfield, G. & Kornhaber, M.L., Eds. Raising Standards or Raising Barriers: Inequality and High-Stakes Testing in Public Education, (pp. 151-178). New York: Century Foundation Press.

Hayward, E. (2001a). "MCAS ratings rankle burbs: Cities produce bigger performance boost," Boston Herald, 10 January: http://www.bostonherald.com/news/local_regional/mcas01102001.htm

Hayward, E. (2001b). "Phys ed teachers say programs reduced," Boston Herald, 29 January: http://www.bostonherald.com/news/local_regional/physed01292001.htm

Heubert, J. P. & Hauser, R. M. Eds. (1999). High Stakes: Testing for Tracking, Promotion, and Graduation. Washington, DC: National Academy Press.

Hoboth, J. (2000). "Monson retools school curriculum," Springfield Union-News, 13 September: http://www.masslive.com/newsindex/metroeast/index.ssf?/news/pstories/m913mcas.html

Hoff, D. J. (2000). "Testing's Ups and Downs Predictable," Education Week, 19(20), 26 January: 1, 12-13.

Huang, J. (2001). "Citizen task force targets test scores," Springfield Union News, 25 January: http://www.masslive.com/newsindex/springfield/index.ssf?/news/pstories/se125sch.html.

Jones, K. & Whitford, B. L. (December, 1997). Kentucky's Conflicting Reform Principles: High Stakes School Accountability and Student Performance Assessment. Phi Delta Kappan, 97 (4): 276-281.

Jones, L. V. (2001). Assessing Achievement Versus High-stakes Testing. A Crucial Contrast. Educational Assessment, 7 (1): 21-28.

Kane, T. J. & Staiger, D. O. (March 2001). "Improving School Accountability Measures," National Bureau of Economic Research, Working Paper No. 8156.

Kane, T. J. & Staiger, D. O. (2001). "Rigid Rules Will Damage Schools," New York Times, 13 August: A21 (also: http://www.nytimes.com/2001/08/13/opinion/13KANE.html).

Kane, T. J. & Staiger, D. O. (Forthcoming, 2002). "Volatility in School Test Scores: Implications for Test-Based Accountability Systems," forthcoming in Diane Ravitch (ed.) Brookings Papers on Education Policy, 2002. Washington, DC: Brookings Institution.

Kane, T. J., Staiger, D. O., and Geppert, J. (2001). "Assessing the Definition of 'Adequate Yearly Progress' in the House and Senate Education Bills." Unpublished paper. 15 July.

Kohn, A. (1999). The Schools Our Children Deserve: Moving Beyond Traditional Classrooms and "Tougher Standards." Boston: Houghton Mifflin.

Kohn, A. (2001). "Emphasis on Testing Leads to Sacrifices in Other Areas," USA Today, 22 August: 15A.

Koretz, D. (1988). "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?" American Educator, Summer: 8-15, 46-52.

Koretz, D.M., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991). The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalizations Across Tests. Presented in R. L. Linn (Chair), Effects of High-Stakes Educational Testing on Instruction and Achievement, Symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, 5 April.

Lipman, L. (2001). "Lowell High School's 9:25 Lunch Could Violate Federal Law," Associated Press, 4 September: http://www.boston.com/dailynews/247/region/Lowell_High_School_s_9_25_a_m_:.shtml.

Madaus, G. & Clarke, M. (2001). The Adverse Impact of High-Stakes Testing on Minority Students: Evidence from One Hundred Years of Test Data. In Orfield, G. & Kornhaber, M.L., Eds. Raising Standards or Raising Barriers: Inequality and High-Stakes Testing in Public Education, (pp. 85-106). New York: Century Foundation Press.

Madaus, G. (1983). Test Scores: What Do They Really Mean in Educational Policy. Presentation to the 1983 Convention of the New Jersey Education Association.

Malley, C. (2001). "Swift emphasized education, calls for prudence on budget," Springfield Union-News, 8 September: http://www.masslive.com/chicopee/unionnews/index.ssf?/news/pstories/ae98swif.html.

Massachusetts Coalition for Authentic Reform in Education. (ND). A Call for An Authentic State-Wide Assessment System. http://www.caremass.org/statements/authentic_statewide%20.htm.

Massachusetts Department of Education. (NDa). Accountability and Targeted Assistance School Performance Rating Process Overview: http://www.doe.mass.edu/ata/spr.html.

Massachusetts Department of Education. (NDb). First Group of Compass Schools Selected to Serve During 2001-2002 School Year. http://www.doe.mass.edu/ata/eval01/prreports/compass/01comp.html.

Massachusetts Department of Education. (NDc). List of Schools Invited to Apply for Exemplary Schools Program, http://www.doe.mass.edu/news/archive01/list0209.html.

Massachusetts Department of Education. (NDd). Report of the School Panel Review of the Reay E. Sterling Middle School, Quincy, MA: http://www.doe.mass.edu/ata/eval01/prreports/compass/sterling.html.

Massachusetts Department of Education (NDe). Report of the School Panel Review of the Saltonstall School, Salem, MA: http://www.doe.mass.edu/ata/eval01/prreports/compass/salton.html

Massachusetts Department of Education. (1999). Press release: "Five principals recognized for MCAS improvement," 22 December: http://www.doe.mass.edu/news/archive99/Dec99/122299pr.html.

Massachusetts Department of Education. (November 21, 2000). Spring 2000 MCAS Tests: Report of 1998-2000 School Results. Malden, MA: Massachusetts Department of Education.

Massachusetts Department of Education. (November 2001a). Spring 2001 MCAS Tests: Spring 2001 MCAS Tests: Report of 2000-2001 School Results. Malden, MA: Massachusetts Department of Education.

Massachusetts Department of Education. (November 2001b). Dropout Rates in Massachusetts Public Schools: 1999-00. Malden, MA: Massachusetts Department of Education.

Massachusetts Department of Education. (2000). Commissioner's Updates, 20 December: http://www.doe.mass.edu/mailings/2000/1220/cm122000.html.

Massachusetts Department of Education. (20 December 2001). Commissioner's Update, "Commonwealth Compass Schools Hold Informational Event:" http://www.doe.mass.edu/ata/news01/csis_event.html.

Massachusetts Department of Education. (August 2001). Commissioner's Special Update to Superintendents and HS Principals about Project Success: The Class of 2003: http://www.doe.mass.edu/news/archive01/su083101.html.

Massachusetts Department of Education. (2001a). Dropout rates in Massachusetts Public Schools, 1999-2000. Malden, MA: Massachusetts Department of Education.

Massachusetts Department of Education. (2001b). Press release: "10 Schools praised for improvement on the MCAS exams," 15 November: http://www.doe.mass.edu/news/news.asp?id=435.

Massey, J. (2000). "New test prep program seeks to boost scores," New Bedford Standard-Times, 31 May: http://www.s-t.com/daily/05-00/05-31-00/a07ed041.htm

Massey, J. (2001a). "NBHS turns to Web for MCAS help," New Bedford Standard-Times, 8 January: http://www.s-t.com/daily/01-01/01-08-01/a01lo003.htm.

Massey, J. (2001). "South Coast schools rethink methods to teaching English," New Bedford Standard-Times, 14 January: http://www.s-t.com/daily/01-01/01-14-01/a01lo008.htm.

McElhenny, J. (2001a). "As tenth graders are tested, more ninth graders being held back," New Bedford Standard-Times, 15 November: http://www.s-t.com/daily/11-01/11-15-01/a03sr016.htm

McElhenny, J. (2001b). "Scoring errors cited in MCAS report," New Bedford Standard Times, 12 January: http://www.s-t.com/daily/01-01/01-12-01/a03sr018.htm.

McFarlane, C. (2000). "MCAS termed 'absolutely critical,'" Worcester Telegram, 22 September: http://www.telegram.com/news/page_one/tests.html.

McGill-Franzen, A. & Allington, R.L. (1993). "Flunk 'em or get them classified: The contamination of primary grade accountability data," Educational Researcher, 22(1): 19-22.

McNeil. L. & Valenzuela, A. (2001). The Harmful Impact of the TAAS System of Testing in Texas: Beneath the Accountability Rhetoric. In Orfield, G. & Kornhaber, M.L., Eds. Raising Standards or Raising Barriers: Inequality and High-Stakes Testing in Public Education, (pp. 127-150). New York: Century Foundation Press.

Miller, N. (2001). "Hudson, Framingham and Natick among few in state to perform above average on MCAS improvement," Metrowest Daily News, 11 January: http://www.townonline.com/metrowest/natick/news/989888_0_hudson__011101_e3ad83ffc5.html.

Moore, K. (2001) State ratings criticized as 'bad math,'" Springfield Union News, 22 January: http://www.masslive.com/newsindex/springfield/index.ssf?/news/pstories/se122mchtml.

Myers, K.C. (2001a). "Crunch time: As test time draws near, teachers try to balance their daily curriculum with MCAS review," Cape Cod Times, 1 April: http://www.capecodonline.com/cctimes/crunchtime1.htm.

Myers, K.C. (2001b). "The remediation equation: Schools help Cape students cram for MCAS retest," Cape Cod Times, 3 December: http://w.ww.capecodonline.com/cctimes/archives/2001/dec/3/theremediation3.htm.

Neill, M. & Gayler, K. (2001). Do High-Stakes Graduation Tests Improve Learning Outcomes?: Using State-level NAEP Data to Evaluation the Effects of Mandatory Graduation Tests. In Orfield, G. & Kornhaber, M.L., Eds. Raising Standards or Raising Barriers: Inequality and High-Stakes Testing in Public Education, (pp. 107-125). New York: Century Foundation Press.

Nichols, M. (2001). "Schools preparing kids for MCAS; vacation schools helps those at risk," Town Online/Community Newspapers, Northwest, 21 February: http://www.townonline.com/northwest/06109784.htm

Oakes, J. (1989). What Educational Indicators?: The Case for Assessing the School Context. Educational Evaluation and Policy Analysis, 11 (2): 181-189

Olson, L. (2001). "Study Questions Reliability of Single-Year Test-Score Gains," Education Week, 20(37), 23 May: 9.

Palm Beach Post. (2001). "FCAT's funny math," Palm Beach Post, 9 August: http://www.gopbi.com/partners/pbpost/epaper/editions/today/opinion_3.html.

Pappano, L. (2001). "MCAS scores gaining power as public labels." Boston Sunday Globe, 21 October: F13.

Paris. S. (2000). Trojan horse in the schoolyard: The hidden threats in high-stakes testing. Issues in Education, 6(1,2): 1-16.

Paris, S.G., & Urdan, T. (2000). Policies and practices of high-stakes testing that influence teachers and schools. Issues in Education, 6(1,2): 83-107.

Perlman, H. (2000). "Principals from schools with most improved MCAS scores honored," Associated Press, 18 December: http://www.boston.com/dailynews/353/region/Principals_from_schools_with_m:.shtml.

Rhode Island Department of Education. (ND). SALT: School Accountability for Learning and Teaching: Frequently Asked Questions About SALT Visits and Reports, http://www.ridoe.net/schoolimprove/salt/faqs.htm.

Scarlett, S. (2001). "Some feel the bell rings a bit too early for lunch at Lowell High," Lowell Sun, 5 September: http://www.lowellsun.com/S-ASP-BIN/REF/Index.ASP?PUID=1697&indx-1071861.

Rene, S. (2001). "Some see short shrift for black history at NBHS," New Bedford Standard-Times, 26 February: http://www.s-t.com/daily/02-01/02-26-01/a09lo044.htm.

Sheldon, K. M. & Biddle, B. J. (1998). "Standards, Accountability, and School Reform: Perils and Pitfalls," Teachers College Record, 100 (1), Fall: 164-180.

Sirotnik, K. A. and Kimball, K. (1999). Standards for Standards-Based Accountability Systems. Phi Delta Kappan, 81(3): 209-214.

Smith, M. L. & Rottenberg, C. (1991). "Unintended Consequences of External Testing in Elementary Schools," Educational Measurement: Issues and Practice, 10(4), Winter: 7-11.

Smith, M. L. & Shepard, L. (Eds.) (1989). Flunking Grades: Research and Policies on Retention. New York: Falmer Press.

Socolar, P. (2001). "State performance awards: few schools repeat as winners," Philadelphia Public Schools Notebook, 8(2), Winter 2000-01:20.

Socolar, P. (2001). Personal communication, 15 July.

Stecher, V. & Barron, S. (1999). "Test-Based Accountability: The Perverse Consequences of Milepost Testing." Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 1999.

Sutner, S. & McFarlane, C. (2001). "Rating system faulted," Worcester Telegram, 10 January: http://www.telegram.com/news/page_one/mcas1.html.

Tantraphol, R. (2001). "Ratings add to MCAS test controversy." Springfield Union News, 10 January: http://www.masslive.com/newsindex/index.ssf?/news/stories/tn110rat.html.

Tomei, F. (2001). "State MCAS grade raises super's ire, Ipswich Chronicle, 11 January: http://www.townonline.com/north/ipswich/news/989888_0_state_011101_2936e753e2.html

Tuerck, D.G. (2001). "MCAS rating system needs to be fixed," Boston Globe, 20 January: A15.

Vaishnav, A. (2001). Some top-scoring schools faulted; many question MCAS assessment. Boston Globe, 10 January: A01.

Walsh, M. (2000). "More Incentives Would Drive Schools To Improve, Business Alliance Argues," Education Week, 19 (23), 16 February: 8.

Wehlage, G. & Rutter, R. (1986). Dropping out: How much do schools contribute to the problem? Teachers College Record 87(3). Spring: 374-392.

Wheelock, A. (1998). Safe To Be Smart: Building a Culture for Standards-Based Reform in the Middle Grades. Columbus, OH: National Middle School Association.

Whitford, B.L. & Jones, K. (2000). Accountability, Assessment, and Teacher Commitment: Lessons from Kentucky's Reform Efforts. Albany: State University of New York Press.

Wilson, C. (2001). "Challenges for New Superintendents," Daily Hampshire Gazette, 22 August: http://www.gazettenet.com/08222001/schools/5592.htm.

Wicka, N. (2000). Summer session hones test skills, Daily Hampshire Gazette, 4 July: http://www.gazettenet.com/07042000/schools/26964.htm.

Wilson, T. A. (1996). Reaching for a Better Standard: English School Inspection and the Dilemma of Accountability for American Public Schools. New York: Teachers College Press.